8 Machine Learning and Real-world Data (sht25)

Two systems are classifying documents into relevant to a particular topic (e.g., "remedies against hair loss"), or not. System A is given 5 documents. It classified all as relevant, but only 4 of these are relevant according to a human. System B is given 20 documents. It classified all as relevant, but only 5 are indeed relevant.

(a) Define precision, recall and accuracy and calculate these for Systems A and B. [4 marks]

(b) You now find out that the situation was actually quite different: The systems were asked to make relevance decisions on documents from a large pool (1 million documents), and System A classified the 5 documents mentioned above from the document set as relevant, and the rest as irrelevant; in a similar way, System B classified the 20 documents mentioned above as relevant. You also learn that this topic had 6 relevant documents in total. With this information, how do your measurements of precision, recall and accuracy change? [2 marks]

(c) In a situation such as the one in Part (b), which of the three evaluation metrics accuracy, precision and recall are appropriate, and why? [2 marks]

(d) Consider now the four systems A, B, C, D in the table below, which ranked documents according to relevance, for the topic introduced in Part (b). The table shows the top 20 ranks. X stands for "relevant document".

	Rank																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	–	–	X	–	–	X	–	–	–	X	X	–	–	–	–	–	–	–	X	–
B	X	X	–	X	–	–	–	X	–	–	–	–	–	–	–	–	–	X	–	X
C	–	–	–	–	X	–	–	–	–	–	–	–	X	X	–	X	–	–	X	X
D	X	X	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	X	X	X

Precision at rank 10 ($P_{Rank10}(A)$) and recall at rank 10 ($R_{Rank10}(A)$) are two metrics which are calculated at rank 10. Calculate these for systems A–D. [3 marks]

(e) Two new metrics are proposed: $Precision@Recall = 0.5$, abbreviated $P@R = 0.5$, is defined to be the precision at the rank where recall first reaches 50%; $Recall@Precision = 0.5$, abbreviated $R@P = 0.5$, is defined to be the rank where precision first reaches 50%. Calculate these metrics for the systems A–D. (Tip: it is not guaranteed that these measurements can be taken in all cases.) [4 marks]

(f) What is problematic about the metrics from Part (d), and what is problematic about the metrics from Part (e)? Which properties would an ideal metric for measuring classification performance in the context of extremely imbalanced tasks for ranked systems have? [5 marks]